

The Effect of Recurrent Mutation on the Linkage Disequilibrium under a Selective Sweep

C. Borck*

September 17, 2010

Abstract

A selective sweep describes the reduction of diversity due to strong positive selection. If the mutation rate to a selectively beneficial allele is sufficiently high, Pennings and Hermisson (2006a) have shown, that it becomes likely, that a selective sweep is caused by several individuals. Such an event is called a soft sweep and the complementary event of a single origin of the beneficial allele, the classical case, a hard sweep. We give analytical expressions for the linkage disequilibrium (LD) between two neutral loci linked to the selected locus, depending on the recurrent mutation to the beneficial allele, measured by D and $\widehat{\sigma_D^2}$, a quantity introduced by Ohta and Kimura (1969), and conclude that the LD-pattern of a soft sweep differs substantially from that of a hard sweep due to haplotype structure. We compare our results with simulations.

1 Introduction

It is a long-standing question of evolutionary biology to decide about the relative importance of evolutionary factors such as selection versus genetic drift to shape patterns of diversity. Today this topic is studied based on DNA variation data taken from a sample of a population. Important work on the effects of positive selection on patterns in DNA data was made by Maynard Smith and Haigh (1974). They showed, that neutral variation linked to a beneficial allele also increases in frequency. This is called the hitchhiking effect and the resulting reduction of neutral variation is termed a selective sweep. When a beneficial allele fixes in a population, this allele can have a single or several origins, i.e. it can be brought to the population by a single or by several mutants. If several individuals, called founders, account for the fixation of a beneficial allele, we will talk as Pennings and Hermisson about a soft sweep and else about a hard sweep.

There are various reasons for a soft sweep. Adaptation can occur from recurrent migration, mutation or act on standing genetic variation. We treat here the case of recurrent mutation, which also applies to migration in a special case. Realistic models for recurrent migration in general can lead to more complex scenarios due to population structure.

It has been shown by Hermisson and Pennings in (Hermisson and Pennings, 2005; Pennings and Hermisson, 2006a), that soft sweep events become frequent, if the scaled mutation rate $\theta_s = 4Nu_s$ (where N is the diploid population size and u_s the mutation probability to the beneficial allele per individual per generation) is sufficiently high. While hard sweeps dominate for $\theta_s < 0.01$, both hard sweeps and soft sweeps occur in the range $0.01 < \theta_s < 1$. For $\theta_s > 1$ almost all adaptive substitutions will result in soft sweeps. Soft sweeps become likely for populations with large population sizes N or for alleles with high recurrent mutation rates u_s . For example, most pathogens have extremely high population sizes. Therefore their genomes are good candidates for the detection of soft selective sweeps, see e.g. (Nair et al., 2007) for research on soft sweeps in malaria parasites. Karasov et al. (2010) concluded lately that in *Drosophila melanogaster* there

*Albert-Ludwigs-University, Freiburg, Eckerstraße 1, 79104 Freiburg, Deutschland, e-mail: Cornelia.Borck@stochastik.uni-freiburg.de

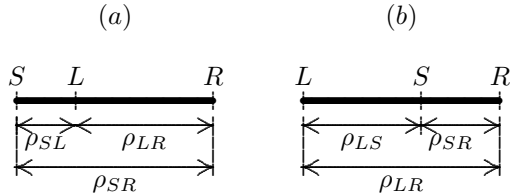


Figure 1: The two possible geometries of the selected locus (S) and the two neutral loci (L and R). The scaled recombination rates between the two loci are given by ρ_{SL} , ρ_{LR} , ρ_{LS} and ρ_{SR} .

should exist a huge amount of soft sweeps due to tremendous short-term effective population sizes relevant for adaptation. Schlenke and Begun (2005) located some of these regions. Recent research by Scheinfeldt et al. (2009) shows, that the DNA pattern around the human gene *ALMS1*, causing the Alstroem Syndrome which presents with early childhood obesity and insulin resistance leading to Type 2 diabetes, may also be the result of a soft sweep. Further Tishkoff et al. (2007) found out, that different SNPs in the human genome all lying in the same short genome region of 110 bp are responsible for the human lactase persistence in the African and European human populations. In their studies of LD (measured by the D' value and the LOD score) the pattern of a soft sweep can be recognized. Ongoing research argues for the importance of soft sweeps and polygenic adaptation, see for a review about this discussion e.g. (Pritchard et al., 2010).

In order to detect soft sweeps it is important to understand the footprints they leave in DNA data. For this purpose it is necessary to make statistical predictions available, which allow us to find targets of recent positive selection. Pennings and Hermisson (2006a) showed that tests based on haplotype structure have high power to detect soft sweeps. If a soft sweep occurred, the population can be divided into several haplotype groups, one for each founder. Without mutation and recombination during the sweep the genomes of the groups differ at the same loci as the founders differed in the beginning of the sweep. Especially, in the case of two founders each allelic variant of a SNP locus is always linked to a single haplotype group. So high linkage disequilibrium of two neutral loci in a neighborhood of the selected locus should be found. This gives rise to the conjecture, that linkage disequilibrium is a useful quantity to detect soft sweeps.

LD has been computed under neutrality by Ohta and Kimura (1969). Stephan et al. (2006), McVean (2007) and Pfaffelhuber et al. (2008) gave analytical expressions for measures of LD after a hard selective sweep. Kim and Nielsen (2004) developed a composite-likelihood method for detecting hard sweeps incorporating information from measures of linkage disequilibrium based on simulation studies. The aim of this article is to give analytical expressions for linkage disequilibrium under a selective sweep with recurrent mutation to the beneficial allele (see Theorem 3.1 and Theorem 3.2). To determine the linkage disequilibrium we use an extended star-like approximation for the genealogy of the selected site, see Section 2.4. A similar approach was applied by Pfaffelhuber et al. (2008) to obtain the linkage disequilibrium in the case of a hard sweep. We will see, that soft sweeps produce a different signal than hard sweeps. In Section 4 we compare our computations with simulations.

2 Model and measures of linkage disequilibrium

2.1 The frequency of the selected locus during a sweep

We consider a DNA region of a population of N diploid individuals and concentrate on the neighborhood of a bi-allelic selected locus S with a wild-type allele b and beneficial allele B . The new beneficial allele B with fitness advantage s enters the population recurrently by mutation and is assumed to fix eventually. The population reproduces at the beneficial locus according to the Moran model in continuous time with selection and recurrent mutation to the beneficial allele,

i.e. denoting by $(X^N(t))_{t \geq 0}$ the frequency of the individuals carrying the beneficial allele in a population of size N , $(X^N(t))_{t \geq 0}$ is a jump Markov process with transition rates from

$$i/2N \text{ to } (i+1)/2N \text{ at rate } u_s(2N-i) + (2N-i)\frac{i}{2N}(1+s) \quad (1)$$

$$i/2N \text{ to } (i-1)/2N \text{ at rate } i\frac{(2N-i)}{2N}, \quad (2)$$

with $u_s, s \geq 0$. The rate $u_s(2N-i)$ is the mutation rate, the rates $(2N-i)\frac{i}{2N}(1+s)$, $i\frac{(2N-i)}{2N}$, respectively, are resampling rates which change the frequency of the beneficial allele by plus $\frac{1}{2N}$, minus $\frac{1}{2N}$, respectively. Of course, resampling events inside the beneficial and wild-type locus, which do not change the frequency of the beneficial allele, are also possible.

The frequency of the beneficial allele can be approximated for large N by a differential equation:

Proposition 2.1. *Denote by $X^N(t)$ the frequency of the beneficial allele in the Moran model with constant diploid population size N at time t . Let $X^N(0) = \frac{\lceil \epsilon 2N \rceil}{2N}$. Then the frequency of the beneficial allele $X^N(t)$ converges for $N \rightarrow \infty$ to the solution of the differential equation:*

$$\dot{X}(t) = u_s(1 - X(t)) + sX(t)(1 - X(t)) \quad (3)$$

with initial condition

$$X(0) = \epsilon,$$

in the sense, that for all $\delta > 0$ and all t

$$\lim_{N \rightarrow \infty} P(\sup_{s \leq t} |X^N(s) - X(s)| > \delta) = 0.$$

The proof is an easy application of Theorem 3.1 in (Kurtz, 1971).

Equation (3) has the solution

$$X(t) = \frac{(\epsilon s + u_s)se^{st} - (s - \epsilon s)u_s e^{-u_s t}}{s((\epsilon s + u_s)e^{st} + (s - \epsilon s)e^{-u_s t})}. \quad (4)$$

In this approximation we say that the allele fixes in the population, if $X_T = 1 - \epsilon$. For the above equation, this happens at time $T = \frac{1}{s+u_s} \log \left(\frac{(1-\epsilon)(u_s+s(1-\epsilon))}{\epsilon(\epsilon s + u_s)} \right)$. With $\epsilon = 1/\alpha$, we obtain $T = \frac{1}{s+u_s} \log \left(\frac{(\alpha-1)(\alpha-1+(\theta_s/2))}{1+(\theta_s/2)} \right)$, denoting by $\alpha := 2Ns$ and $\theta_s := 4Nu_s$. For small $u_s \ll s$ and large N , such that $\alpha \gg 1$, the fixation time T is approximately $2(\log \alpha)/s$. The fixation time will be relevant below.

2.2 Measures of linkage disequilibrium

Our aim is to provide analytical results for the linkage disequilibrium of two neutral loci in a neighborhood of the selected locus. Different quantities have been proposed to measure the association of two loci. We will compute two of them here. Consider two neutral loci L and R , linked to the selective locus S . The neutral loci can either lie both on the same side of the selected locus or the selected locus lies between the neutral loci, see Figure 1. (If both neutral loci lie on the left side of the selected locus, we name the leftmost locus R -locus and the locus in the middle L -locus, i.e. we have the ordering $R L S$.) We consider only loci with exactly two allelic variants. Denote them by L/ℓ and R/r and their allelic frequencies by q_ℓ, q_L , etc.

Definition 2.2 ($D_{\ell,r}$ and $\widehat{D}_{\ell,r}$). *The simplest approach to measure linkage disequilibrium between the allele ℓ of the L -locus and the allele r of the R -locus is to compute the quantity*

$$D_{\ell,r} := q_{\ell r} - q_\ell q_r. \quad (5)$$

If $D_{\ell,r}$ is zero the alleles ℓ and r are said to be in linkage equilibrium, else in linkage disequilibrium.

In practice, the population frequencies q_ℓ, q_r , etc. are often not available, but only the allelic frequencies in a sample, $\widehat{q}_\ell, \widehat{q}_r$, etc. In samples LD can be measured by

$$\widehat{D}_{\ell,r} := \widehat{q}_{\ell r} - \widehat{q}_\ell \widehat{q}_r.$$

Remark 2.3. An easy calculation shows, that

$$D_{\ell,r} = D_{L,R} = -D_{\ell,R} = -D_{L,r}$$

and analogous equalities hold for \widehat{D} .

Averaging $D_{\ell,r}$ over all allelic variants gives zero due to Remark 2.3. Hence it makes sense to consider $D_{\ell,r}^2$. Since $D_{\ell,r}^2 = D_{\ell,R}^2 = D_{L,R}^2 = D_{L,r}^2$ the quantity $D_{\ell,r}^2$ actually does not depend on the allelic variant ℓ, r . Therefore we write D^2 instead. However D^2 depends strongly on the size of the allelic variants: If q_ℓ and q_r are small, D^2 is also small, whenever the allelic variants may be not in association at all. Therefore the so called standard linkage disequilibrium, introduced by Ohta and Kimura (1969), is often considered.

Definition 2.4 (σ_D^2 and $\widehat{\sigma}_D^2$). The standard linkage disequilibrium $\sigma_D^2, \widehat{\sigma}_D^2$ in the sample, respectively, is given by

$$\sigma_D^2 = \frac{\mathbb{E}[D_{\ell,r}^2]}{\mathbb{E}[q_\ell(1-q_\ell)q_r(1-q_r)]},$$

and

$$\widehat{\sigma}_D^2 = \frac{\mathbb{E}[\widehat{D}_{\ell,r}^2]}{\mathbb{E}[\widehat{q}_\ell(1-\widehat{q}_\ell)\widehat{q}_r(1-\widehat{q}_r)]},$$

respectively.

Remarks 2.5.

- Note, that $\frac{\mathbb{E}[\widehat{D}_{\ell,r}^2]}{\mathbb{E}[\widehat{q}_\ell(1-\widehat{q}_\ell)\widehat{q}_r(1-\widehat{q}_r)]}$ does not depend on the particular allelic variant ℓ, r , too. Therefore it makes sense to write $\widehat{\sigma}_D^2$, instead of $\widehat{\sigma}_{D_{\ell,r}}^2$.
- We compute linkage disequilibrium during the sweep. So, if time is important, we will write $D_{\ell,r}(t) := q_{\ell r}(t) - q_\ell(t)q_r(t)$, etc.
- Pfaffelhuber et al. (2008) have computed $\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T)]$ and $\widehat{\sigma}_D^2$ for a hard sweep, i.e. the case $u_s = 0$. See Figure 6 for a plot of $\widehat{\sigma}_D^2$ under neutrality and for $\theta_s = 0$ and $\theta_s = 0.1$.
- Naturally one would consider the quantity

$$r^2 := \mathbb{E}\left[\frac{\widehat{D}^2}{\widehat{q}_L(1-\widehat{q}_L)\widehat{q}_R(1-\widehat{q}_R)}\right].$$

But this quantity is less attractive for analytical studies, because it is mathematically difficult to handle. However, see the recent paper of Song and Song (2007) for an analytical computation of r^2 under neutrality.

2.3 Genealogies: Motivation

We want to compute $\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T)]$ and $\widehat{\sigma}_D^2$ at the end of the sweep assuming small sample sizes $n \ll N$ for the computation of $\widehat{\sigma}_D^2$. We will use the 1-1-correspondence between the probability to draw two pairs of heterozygous neutral loci and $\widehat{\sigma}_D^2$, (see step 3 of the proof of Theorem 3.2). The probability to draw a heterozygous pair at the end of the sweep differs from the probability to

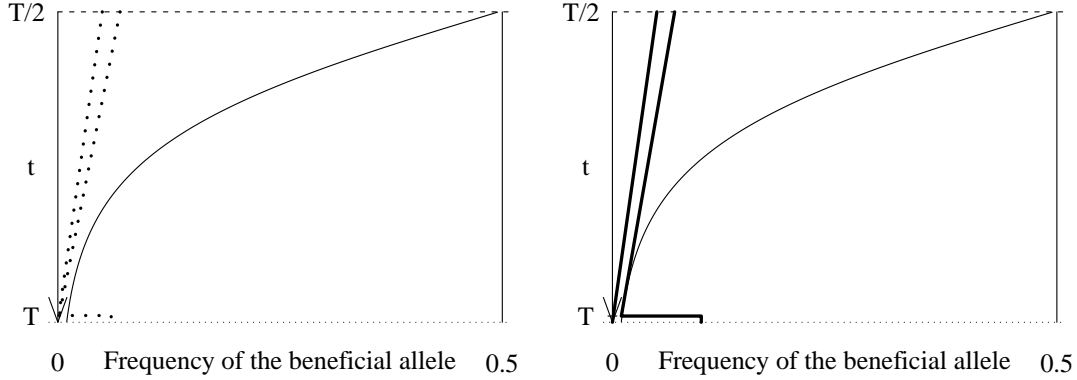


Figure 2: The figure on the left shows two lines which coalesce before they mutate. The figure on the right shows two lines which mutate before they coalesce (further back in the past).

draw a heterozygous pair at the beginning of the sweep due to the change of the genealogy during the sweep. We shall start with heterozygous pairs of a sample taken from the population at the end of the sweep and follow the lines of the pairs till the beginning of the sweep. In our notation time is running backwards starting from time T of fixation, i.e. if $t_2 > t_1$ the time t_2 lies further back in the past than the time t_1 . For example $\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T)]$ is the expected value of $D_{\ell,r}$ at the end of the sweep given $D_{\ell,r}$ at the beginning of the sweep.

To define the genealogies of two neutral loci in the neighborhood of a selected locus in a Moran model we would have to extend the Moran model as introduced in Section 2.1 to a full three-locus model. However, multi-locus genealogies under such a Moran model are very complex. Under certain conditions star-like genealogies approximate the genealogies of the Moran model quite well and allow a computation of the above probabilities due to independent genealogical lines. In the following we introduce such star-like genealogies and justify why it is reasonable to use them in our setting.

We suppose, that neutral mutations occur according to a Poisson Process with rates of order $\mathcal{O}(1/N)$. Since the sweep takes only of order $\log(2Ns)/s$ time units, we can ignore neutral mutations during the sweep. Moreover, back-mutations are rapidly sorted out as they have no fitness advantage. Hence we will ignore back-mutations, too.

Coalescent and mutations to the beneficial allele

- The rate of coalescence of two lines at time t under the condition, that the two lines are in the beneficial background at time t_- and the frequency of beneficial allele is $X_{t_-} = \frac{i}{2N}$ is equal to

$$\frac{\frac{1}{(2N)(2N-1)} \frac{i(i-1)}{2N}}{\frac{i(i-1)}{(2N)(2N-1)}} + \frac{\frac{1}{(2N)(2N-1)} \frac{(i-1)(2N-i+1)(1+s)}{2N}}{\frac{i(i-1)}{(2N)(2N-1)}} = \frac{(1+s)(2N+1)}{2Ni} - \frac{s}{2N}$$

for $i > 1$. The parents of the beneficial offspring are either both from the beneficial background or one is from the wild-type and the second from the beneficial background; for similar calculations see (Barton et al., 2004), Lemma 2.4. For large N this rate is approximately

$$\frac{1+s}{2NX(t)}, \quad (6)$$

since for large N the frequency $X^N(t)$ is well approximated by the solution $X(t)$ of the differential Equation (3) by Proposition 2.1.

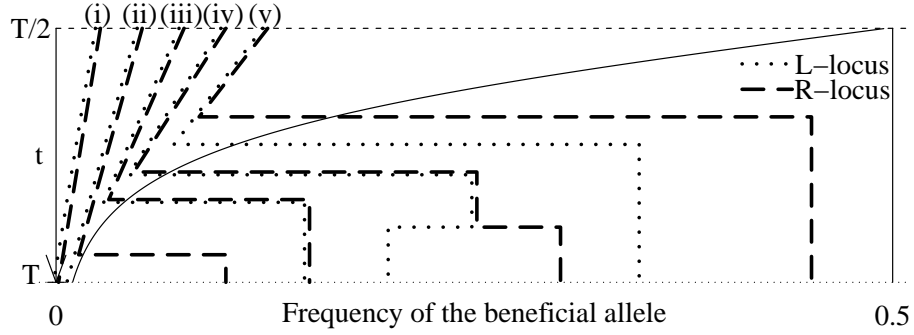


Figure 3: Possible lines in the time interval $[T/2, T]$ (backward in time) for geometry (a). The corresponding probabilities are given in Table 1.

- If an individual mutates to the beneficial type, the genealogical line of this individual jumps (forward in time) from the wild-type background to the beneficial background. Backward in time the line is located at time t_- in the beneficial background and at time t , after the mutation event, in the wild-type background.

The rate of mutation to the beneficial background of a line at time t under the condition, that the frequency of the beneficial allele is $X_{t_-} = \frac{i}{2N}$ at time t_- , is equal to

$$\frac{u_s \frac{2N-i}{2N}}{\frac{i}{2N}},$$

Analogous argumentations as for the coalescence rate yield, that this rate is approximately

$$\frac{u_s(1 - X(t))}{X(t)} \quad (7)$$

for large N .

Both rates scale with $\frac{1}{X(t)}$, which means that the coalescent and mutation rates are high, if the frequency $X(t)$ is small. Hence it makes sense to assume that all mutations to the beneficial allele and all coalescent events occur at time $t = T$, i.e. at the beginning of the sweep. This scaling of the backward mutation rate shows that the star-like approximation, which has before been used for the classical hard sweeps case, should also be appropriate for soft sweeps.

With the approximate mutation and coalescent rates (7) and (6) the probability for a hard sweep in a sample of two can be bounded, see also (Pennings and Hermisson, 2006a) for a similar calculation in a Wright-Fisher-model formulation. The probability for a hard sweep of two lines equals the probability, that the coalescent event happens before the mutation event. The mutation rate in a sample of two lines is approximately $2 \frac{u_s(1-X(t))}{X(t)} = \frac{\theta_s(1-X(t))}{2NX(t)}$, if terms of order $(2Nu_s)^2$ are ignored.

Let $C = (C_t)_{0 \leq t \leq T}$, $M = (M_t)_{0 \leq t \leq T}$ be two independent Poisson processes with rates $\lambda(t) = \frac{1+s}{2NX(T-t)}$, $\mu(t) = \frac{\theta_s(1-X(T-t))}{2NX(T-t)}$ respectively. Denote by S_1 the first jump time of C and by T_1 the first jump time of M . Then the probability for a hard sweep in a sample of two is

$$P_{hard,2} := P(S_1 < T_1).$$

case	event	probability
(i)	no recombination event	$p_{SL}p_{LR}$
(ii)	a LR -recombination event makes the allele at the R -locus escape the sweep without the allele at the L -locus	$p_{SL}(1 - p_{LR})$
(iii)	by a SL -recombination event the line escapes the sweep and the alleles at the L - and R -locus stay linked	$(1 - p_{SL})p_{LR}$
(iv) (v)	a SL -recombination event brings the alleles at the L - and R -loci linked into the wild-type background; here, the ancestry of both alleles is split by a LR -recombination a LR - and a SL -recombination event bring first the allele at the R -locus and then the allele at the L -locus into the wild-type background	$\mathbb{P}[(iv) \text{ or } (v)]$ $= (1 - p_{SL})(1 - p_{LR})$

Table 1: Probabilities of several events happening between times $T/2$ and T_- for geometry (a); see Figure 3. All events are described backwards in time.

where $\theta_s(n) = \theta_s \cdot (\theta_s + 1) \cdots (\theta_s + n - 1)$. See (Pennings and Hermisson, 2006b) for a derivation of this formula in a Wright-Fisher-model formulation.

We will assume in our approximation of the genealogy, that the number of founders and the number of their offspring is Ewens distributed as given in Equation (8).

Recombination events

Forward in time at a recombination event two lines merge into one. If a recombination event occurs between two neighboring loci L_1 and L_2 (we will write L_1L_2 -recombination event, for short), such that L_1 lies on the left side of L_2 , the offspring carries at all loci left of the locus L_1 including the locus L_1 the alleles of the first parent and at the remaining loci the alleles of the second parent (with $L_1, L_2 \in \{L, R, S\}$). Backward in time at a recombination event one line splits up into two lines.

Since coalescence events occur at the beginning of the sweep, one can assume, that each recombination event affects only a single line. The probability for no recombination event in the time interval $[t_1, t_2]$ is given by the probability, that the first jump time of a Poisson process started at time t_1 with rate $r(t)$ does not occur until time t_2 . The rate $r(t)$ depends on the different kinds of recombination events and is specified in the following.

- The frequency X_t stays between backward time 0 and $T/2$ almost the whole time near 1 and is certainly greater $1/2$. (The larger α the longer X_t remains in a small neighborhood of 1.) So, in the first half, recombination between the backgrounds is not frequent. Furthermore if L , R and S are arranged according to geometry (a) SL -recombination events inside the beneficial background cannot be seen in the DNA-data. The only events that can be recognized in the DNA-data and occur at a non negligible amount are LR -recombination events in the beneficial background. If the loci are arranged according to geometry (b) all recombination

line	event	probability
(i)	no recombination event	$p_{LS}p_{SR}$
(ii)	a SR -recombination event makes the allele at the R -locus escape the sweep without the allele at the L -locus	$p_{LS}(1 - p_{SR})$
(iii)	a LS -recombination event makes the allele at the L -locus escape the sweep without the allele at the R -locus	$(1 - p_{LS})p_{SR}$
(iv) (v)	a LS -recombination event followed by a SR -recombination event bring the alleles at the L - and R -locus into the wild-type background same as (iv) but in reverse order of the LS - and SR -recombination events	$\mathbb{P}[(iv) \text{ or } (v)]$ $= (1 - p_{LS})(1 - p_{SR})$

Table 2: Probabilities for events happening between time $T/2$ and T_- for geometry (b); see Figure 4. All events backward in time

events in the beneficial locus may be seen in the data.

The rate of recombination events between loci L_1 and L_2 (with $L_1, L_2 \in \{L, R, S\}$ for geometry (b) and $L_1 = L$ and $L_2 = R$ for geometry (a)) in the beneficial background is approximately $\frac{r_{L_1 L_2} X(T-t) X(T-t)}{X(T-t)}$ with $r_{L_1 L_2} \geq 0$. Therefore the probability for no $L_1 L_2$ -recombination event is given by

$$\exp\left(-\int_0^{T/2} r_{L_1 L_2} X(T-t) dt\right). \quad (9)$$

As long as u_s is small, the differential Equation (3) is only a small perturbation of the differential equation $\dot{X}(t) = sX(t)(1 - X(t))$. For this equation the integral in Equation (9) is equal to $r_{L_1 L_2} \log(\alpha)/s + r_{L_1 L_2} \log(2)/s$. Since the second summand is small, we approximate (9) by

$$p_{L_1 L_2} := \exp(-\rho_{L_1 L_2} \log(\alpha)/\alpha),$$

where $\rho_{L_1 L_2} := 2Nr_{L_1 L_2}$ denotes the recombination rate between the locus L_1 and L_2 .

- In the time interval $[T/2, T]$ all recombination events with offspring in the beneficial background except recombination events inside the beneficial locus are probable to be seen in the data. Similar arguments as above lead to the following assumption: The probability for no recombination between locus L_1 and L_2 in the time interval $[T/2, T]$ is given by

$$p_{L_1 L_2} := \exp(-\rho_{L_1 L_2} \log(\alpha)/\alpha).$$

See Figure 3 - 5 for an illustration of the different types of recombination events possible in the time intervals $[0, T/2)$ and $[T/2, T)$.

With this motivation, we can define an extended star-like genealogy:

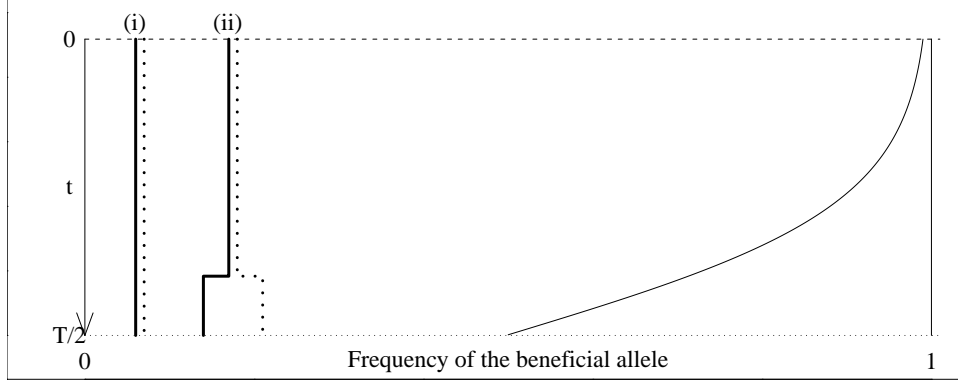


Figure 5: Possible split of two linked neutral loci: Two alleles at the neutral loci linked to the beneficial allele either (i) have a common ancestor at time $T/2$ or (ii) have two different ancestors that are both linked to a beneficial allele.

2.4 Genealogies: Definition

The joint genealogy of two neutral loci in the neighborhood of the selected locus can be defined as a structured partition-valued process. Denote by $\Xi_A := \{\xi_A | \xi_A \text{ partition of } A\}$ the set of partitions of a set A . A partition $\xi_A = \{\xi_1, \dots, \xi_m\}$ is called finer than a partition $\xi'_A = \{\xi'_1, \dots, \xi'_k\}$, iff for each $j \in \{1, \dots, m\}$ exists an $i \in \{1, \dots, k\}$, such that $\xi_j \subseteq \xi'_i$. We write $\xi_A \preceq \xi'_A$, if ξ_A is finer than ξ'_A . Let $\xi = \{\xi_1, \dots, \xi_m\} \in \Xi_A$ and $\eta = \{\eta_1, \dots, \eta_k\} \in \Xi_{\{1, \dots, m\}}$, $m \geq k$, then the composition

$$\eta \circ \xi := \left\{ \bigcup_{j \in \eta_1} \xi_j, \dots, \bigcup_{j \in \eta_k} \xi_j \right\}.$$

A structured partition of A is a tuple (ξ_A^B, ξ_A^b) with $\{\xi_A^B \cup \xi_A^b\} \in \Xi_A$ and $\xi_A^B \cap \xi_A^b = \emptyset$. Partition elements in ξ_A^B (ξ_A^b) are called beneficial (wild-type). Denote by

$$\Xi_A^{B,b} := \{(\xi_A^B, \xi_A^b) | (\xi_A^B, \xi_A^b) \text{ is a structured partition of } A\}$$

the set of structured partitions of the set A . Elements of a structured partition (ξ_A^B, ξ_A^b) are of the form (ξ_1, ξ_2) with $\xi_1 \in \xi_A^B$ and $\xi_2 \in \xi_A^b$.

Define $\ell := \{1, \dots, n\}$ the set of the L-loci and $\mathbf{r} := \{n+1, \dots, 2n\}$ the set of the R-loci of a sample of size n of the population. We are interested in the structured partitions of $\ell \cup \mathbf{r}$.

For geometry (a) the different kinds of recombination events can change the structured partition (ξ^B, ξ^b) to

- $(\xi^B \setminus \{\xi_j^B\}, \xi^b \cup \{\xi_j^B\})$, if an *SL*-recombination event happens between a wild-type and beneficial line and the offspring carries at the selected locus the beneficial allele (thus at the *L*- and *R*-locus the individual carries the alleles of the wild-type line).
- $((\xi^B \setminus \{\xi_j^B\}) \cup \{\xi_j^B \cap \ell\}, \xi^b \cup \{\xi_j^B \cap \mathbf{r}\})$, if an *LR*-recombination event happens between an individual of the beneficial background and an individual of the wild-type background and at the *S* and *L*-locus the beneficial line is carried on (forward in time).
- $((\xi^B \setminus \{\xi_j^B\}) \cup \{\xi_j^B \cap \ell, \xi_j^B \cap \mathbf{r}\}, \xi^b)$, if an *LR*-recombination event happens between two individuals of the beneficial background.

- $(\xi^B, (\xi^b \setminus \{\xi_k^b\}) \cup \{\xi_k^b \cap \ell, \xi_k^b \cap \mathbf{r}\})$, if an LR -recombination event happens between two individuals of the wild-type background.
- $(\xi^B \cup \{\xi_k^b\}, \xi^b \setminus \{\xi_k^b\})$, if an SL -recombination event happens between a beneficial and a wild-type line and the offspring carries the beneficial allele.
- $(\xi^B \cup \{\xi_k^b \cap \mathbf{r}\}, (\xi^b \setminus \{\xi_k^b\}) \cup \{\xi_k^b \cap \ell\})$, if an LR -recombination event happens between a beneficial and wild-type line and at the selected and L -locus the wild-type is carried on (forward in time).

For geometry (b) the partitions change in an analogous manner.

Before we give the definition of an extended star-like genealogy we define genealogies and samples.

Definition 2.6. *The genealogy of a set A is a four-time step Markov chain*

$$(\xi_t)_{t \in \{0, T/2, T_-, T\}} = ((\xi_0^B, \xi_0^b), (\xi_{T/2}^B, \xi_{T/2}^b), (\xi_{T-}^B, \xi_{T-}^b), (\xi_T^B, \xi_T^b))$$

with state space $\Xi_A^{B,b}$.

A set $\ell \cup \mathbf{r}$ with $\ell := \{1, \dots, n\}$ and $\mathbf{r} := \{n+1, \dots, 2n\}$, $n \in \mathbb{N}$, is a sample at two loci L and R taken from the population at time $t = 0$, if the genealogy of the sample is at time $t = 0$ given by $\xi_0 = (\{\{1, n+1\}, \dots, \{n, 2n\}\}, \{\emptyset\})$.

With this we can define an extended star-like genealogy as a four-time step random experiment:

Definition 2.7. *An extended star-like genealogy of a sample $\ell \cup \mathbf{r}$ with $\ell := \{1, \dots, n\}$ and $\mathbf{r} := \{n+1, \dots, 2n\}$ at two loci L and R in the neighborhood of a selected locus S arranged according to geometry (a) (resp. geometry (b)) is a four-time step Markov chain*

$$(\xi_t)_{t \in \{0, T/2, T_-, T\}} = ((\xi_0^B, \xi_0^b), (\xi_{T/2}^B, \xi_{T/2}^b), (\xi_{T-}^B, \xi_{T-}^b), (\xi_T^B, \xi_T^b))$$

with state space $\Xi_{\ell \cup \mathbf{r}}^{B,b}$ with the following properties:

- At time $T/2$:
 - Structured partition elements are stochastically independent
 - No recombination events between the backgrounds and no mutations to the beneficial allele, i.e. $P(\xi_{T/2}^b = \emptyset) = 1$
 - No coalescence events, i.e. $P(\xi_{T/2}^B \preceq \xi_0^B) = 1$
 - For geometry (a): No LR -recombination events in the beneficial background with probability p_{LR} , i.e. for $j \in \mathbf{l}$ a structured partition element at time $t = 0$ of the form $(\{\{j, j+n\}\}, \{\emptyset\})$
 - * is kept at time $T/2$ with probability p_{LR}
 - * and changed to $(\{\{j\}, \{j+n\}\}, \{\emptyset\})$ with probability $1 - p_{LR}$
 - For geometry (b): Neither a LS -recombination events nor a SR -recombination events in the beneficial background happens with probability $p_{LS\text{PSR}}$, i.e. for $j \in \mathbf{l}$ a structured partition element at time $t = 0$ of the form $(\{\{j, j+n\}\}, \{\emptyset\})$
 - * is kept till time $T/2$ with probability $p_{LS\text{PSR}}$
 - * and changed to $(\{\{j\}, \{j+n\}\}, \{\emptyset\})$ with probability $1 - p_{LS\text{PSR}}$
- At time T_- :
 - Structured partition elements are stochastically independent

- No coalescence events, i.e. $P(\xi_{T-}^B \cup \xi_{T-}^b \preceq \xi_{T/2}^B) = 1$
- For geometry (a): For $j \in \mathbf{l}$
 - * a partition element at time $T/2$ of the form $(\{j, j+n\}, \{\emptyset\})$ is kept at time T_- with probability $p_{SL}p_{LR}$, changed to $(\{\emptyset\}, \{j, j+n\})$ with probability $(1-p_{SL})p_{LR}$, changed to $(\{\{j\}\}, \{\{j+n\}\})$ with probability $p_{SL}(1-p_{LR})$ and changed to $(\emptyset, \{\{j\}, \{j+n\}\})$ with probability $(1-p_{SL})(1-p_{LR})$.
 - * a partition element at time $T/2$ of the form $(\{j\}, \{\emptyset\})$ is kept at time T_- with probability p_{SL} and changed to $(\{\emptyset\}, \{j\})$ with probability $1-p_{SL}$.
 - * a partition element at time $T/2$ of the form $(\{j+n\}, \{\emptyset\})$ is kept at time T_- with probability p_{SR} and changed to $(\{\emptyset\}, \{j+n\})$ with probability $1-p_{SR}$.
- For geometry (b):
 - * A partition element at time $T/2$ of the form $(\{j, j+n\}, \{\emptyset\})$ is kept at time T_- with probability $p_{LS}p_{SR}$, changed to $(\emptyset, \{\{j\}, \{j+n\}\})$ with probability $(1-p_{LS})(1-p_{SR})$, changed to $(\{\{j\}\}, \{\{j+n\}\})$ with probability $p_{SL}(1-p_{SR})$ and changed to $(\{\{j+n\}\}, \{\{j\}\})$ with probability $(1-p_{SL})p_{SR}$.
 - * A partition element at time $T/2$ of the form $(\{j\}, \{\emptyset\})$ is kept at time T_- with probability p_{LS} and changed to $(\{\emptyset\}, \{j\})$ with probability $1-p_{LS}$.
 - * A partition element at time $T/2$ of the form $(\{j+n\}, \{\emptyset\})$ is kept at time T_- with probability p_{SR} and changed to $(\{\emptyset\}, \{j+n\})$ with probability $1-p_{SR}$.
- At time $t = T$:
At the beginning of the sweep all coalescence and mutation events happen: Let $m \in \mathbb{N}$ and $a_j \in \{0, \dots, m\}$ with $\sum_{j=1}^m ja_j = m$. Denote by $M^m := \{(\xi^B, \xi^b) \in \Xi_{\mathbf{l} \cup \mathbf{r}}^{B,b}; |\xi^B| = m\}$ the set of structured partition of $\mathbf{l} \cup \mathbf{r}$ which beneficial partitions consist of m elements and by $N^{(a_1, \dots, a_m)} := \{\eta = (\eta_1, \dots, \eta_k) \in \Xi_{\{1, \dots, m\}}; \#\{\eta_l; |\eta_l| = j\} = a_j\}$ the set of partitions of $\{1, \dots, m\}$ containing a_j partition elements of size j . Then for $\eta \in N^{(a_1, \dots, a_m)}$

$$P(\xi_T = (\{\emptyset\}, (\eta \circ \xi^B) \cup \xi^b) | \xi_{T-} = (\xi^B, \xi^b) \in M^m) = \frac{m!}{\theta_s(m)} \prod_{j=1}^m \frac{(\theta_s/j)^{a_j}}{a_j!}. \quad (10)$$

We say, that a population evolved according to an extended star-like genealogy, if the genealogy of each sample of the population is extended star-like.

Remark 2.8. If we are interested in the genealogy of a subset M of a sample $\mathbf{l} \cup \mathbf{r}$, the genealogy of M fulfills all conditions of Definition 2.7. In particular, at time $t = T$ the number of founders together with the number of their offspring is Ewens distributed, since Ewens sampling formula is consistent. At time $t = 0$, the genealogy of M is given by

$$\xi_0 = (\{\{1, n+1\} \cap M, \dots, \{n, 2n\} \cap M\}, \{\emptyset\}).$$

In accordance to the possible recombination events during the time interval $[T/2, T)$ we obtain the ancestral lines shown in Figure 3 for geometry (a) and Figure 4 for geometry (b). The probabilities for these events are listed in Table 1 for geometry (a), in Table 2 for geometry (b). For the time interval $[0, T/2)$ the possible ancestral lines are shown in Figure 5. In Figure 2 the left picture shows two lines which coalesce first and mutate then, in the right picture the lines mutate first and coalesce afterwards.

3 Results

Our main result is the computation of the linkage disequilibrium at the end of the sweep measured by $\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T)]$ for two fixed allelic variants ℓ and r and $\widehat{\sigma_D^2}$ for two neutral loci in a neighborhood of the selected locus (backward in time).

We apply the procedure of Pfaffelhuber et al. (2008) to compute $\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T)]$. The main difference between our model and the hard sweep model is, that two lines do not have to coalesce, since both lines may mutate to the beneficial allele.

Theorem 3.1. *Assume, that the population evolved in a DNA-region containing the two neutral loci L and R and the selected locus S according to an extended star-like genealogy and both loci carry exactly two allelic variants ℓ/L and r/R . Then the linkage disequilibrium of the allelic variants ℓ and r measured by $\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T)]$ at the end of the sweep is given by*

$$\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T)] = p_{LR}^2(1 - \frac{1}{1 + \theta_s} p_{SL}^2) D_{\ell,r}(T), \quad (11)$$

if the two neutral loci are arranged according to geometry (a) and

$$\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T)] = p_{LR}^2(1 - \frac{1}{1 + \theta_s}) D_{\ell,r}(T), \quad (12)$$

for L and R arranged according to geometry (b).

Proof. Indeed, consider the genealogy $(\xi)_{t \in \{0, T/2, T_-, T\}}$ of an L -locus $\{1\}$ and an R -locus $\{2\}$, i.e. $\xi_t \in \Xi_{\{1,2\}}^{B,b}$. Denote by d the probability, that the pair $\{1\}, \{2\}$ was linked at the beginning of the sweep, if it is linked at the end of the sweep, i.e. let $d := P(\xi_T = (\{\emptyset\}, \{1, 2\}) | \xi_0 = (\{\{1, 2\}\}, \{\emptyset\}))$. Analogously, denote by e the probability, that the pair has been linked at the beginning, if it is unlinked at the end of the sweep. I.e. $e := P(\xi_T = (\{\emptyset\}, \{\{1, 2\}\}) | \xi_0 = (\{\{1\}, \{2\}\}, \{\emptyset\}))$. Then we can write

$$\begin{aligned} \mathbb{E}[q_{\ell r}(0)|q_{\ell r}(T), q_{\ell}(T), q_r(T)] &= dq_{\ell r}(T) + (1 - d)q_{\ell}(T)q_r(T) \\ \mathbb{E}[q_{\ell}(0)q_r(0)|q_{\ell r}(T), q_{\ell}(T), q_r(T)] &= eq_{\ell r}(T) + (1 - e)q_{\ell}(T)q_r(T) \end{aligned}$$

with $q_{\ell r} \in [0, 1]$ and $q_r, q_{\ell} \in (0, 1)$ and so

$$\mathbb{E}[D_{\ell,r}(0)|D_{\ell,r}(T) = x] = (d - e)D_{\ell,r}(T).$$

The probabilities d and e are for geometry (a) and (b) given by

$$a) \ e = \frac{1}{1 + \theta_s} p_{SL} p_{SR} = \frac{1}{1 + \theta_s} p_{SL}^2 p_{LR} \quad b) \ e = \frac{1}{1 + \theta_s} p_{LS} p_{SR} = \frac{1}{1 + \theta_s} p_{LR}$$

and

$$a) \ d = e(1 - p_{LR}) + p_{LR} p_{LR} \quad b) \ d = e(1 - p_{LR}) + p_{LR} p_{LS} p_{SR} = p_{LR}.$$

In words, a pair is unlinked at the end of the sweep when it was linked at the beginning, iff the pair just coalesces, i.e. neither a recombination event between the S and the L locus neither a recombination event between the S and R locus occurred and the two loci coalesced before they mutated. And a pair which is linked at the end of the sweep is also linked in the beginning, iff either nothing happens or the pair is divided by a LR-recombination event first and then linked again by coalescence event.

From this easily follows Equation (11) for geometry (a) and Equation (12) for geometry (b). \square

To compute the quantity $\widehat{\sigma_D^2}$ consider the three quantities:

$$\begin{aligned} \mathcal{X}_t &:= \mathbb{E}[q_L(t)(1 - q_L(t))q_R(t)(1 - q_R(t))] \\ \mathcal{Y}_t &:= \mathbb{E}[D(t)(1 - 2q_L(t))(1 - 2q_R(t))] \\ \mathcal{Z}_t &:= \mathbb{E}[(D(t))^2] \end{aligned} \quad (13)$$

for $0 \leq t \leq T$.

Theorem 3.2. Given $\mathcal{X}_T, \mathcal{Y}_T$ and \mathcal{Z}_T at the beginning of the sweep and a sample of size n of a population at the end of the sweep, i.e. a set of L -loci $\ell := \{1, \dots, n\}$ and a set of R -loci $\mathbf{r} := \{n+1, \dots, 2n\}$, assume that the genealogy of the sample is extended star-like. Then the standard linkage disequilibrium $\widehat{\sigma}_D^2$ of this sample of two neutral loci at the end of a sweep equals:

$$\widehat{\sigma}_D^2 = \widehat{\mathcal{Z}}_0 / \widehat{\mathcal{X}}_0. \quad (14)$$

with

$$\begin{aligned} \widehat{\mathcal{Z}}_0 = & p_{LR}^4 (p_{SL} - 1)^2 (p_{SL}^2 (\mathcal{X}_T + \mathcal{Y}_T) + (1 + 2p_{SL}) \mathcal{Z}_T) \\ & + \theta_s p_{SR}^2 \left(\frac{\mathcal{X}_T}{3} (p_{LR} (11p_{SR} - 2 - 6p_{LR}) + p_{SR} (2 - 4p_{SR})) \right. \\ & \left. + \frac{\mathcal{Y}_T}{12} (p_{SR} (2 - 21p_{SR}) + p_{LR} (38p_{SR} - 15p_{LR} - 2)) + \frac{\mathcal{Z}_T}{3} (9p_{LR}^2 - 9p_{SR}p_{LR} + p_{SR}^2) \right) \end{aligned} \quad (15)$$

and

$$\begin{aligned} \widehat{\mathcal{X}}_0 = & (1 - p_{SR}) (p_{SL} - 1) (\mathcal{X}_T (1 + p_{SR} + p_{SL}) + (\mathcal{X}_T + \mathcal{Y}_T) (p_{SL} p_{SR})) \\ & + \frac{\theta_s p_{SL}^2}{3} \left(\mathcal{X}_T (3p_{LR}^2 - 5p_{LR} + 3 + 2p_{SR}p_{LR} + 2p_{SR} - 4p_{SR}^2) \right. \\ & \left. + 5\mathcal{Y}_T (p_{SR}p_{LR} + p_{SR} - \frac{17}{20}p_{LR}) + p_{SR}^2 (\mathcal{Z}_T - \frac{21}{4}\mathcal{Y}_T) \right) \end{aligned} \quad (16)$$

for geometry (a) and with

$$\begin{aligned} \widehat{\mathcal{Z}}_0 = & \theta_s p_{LR}^3 \left(\frac{\mathcal{X}_T}{3} (1 - p_{LS} + 2p_{LR} - p_{SR}) \right. \\ & \left. + \frac{\mathcal{Y}_T}{12} (3p_{LR} - p_{SR} - p_{LS} + 1) + \frac{\mathcal{Z}_T}{3} p_{LR} \right) \end{aligned} \quad (17)$$

and

$$\begin{aligned} \widehat{\mathcal{X}}_0 = & \mathcal{X}_T ((1 - p_{LS}^2)(1 - p_{SR}^2)) + \mathcal{Y}_T (p_{LR}(1 - p_{LS})(1 - p_{SR})) \\ & + \theta_s \left(\frac{\mathcal{X}_T}{3} (3p_{LS}^2 + 2p_{LR}p_{SR} - 4p_{LR}^2 - 5p_{LR} + 2p_{LR}p_{LS} + 3p_{SR}^2) \right. \\ & \left. + \frac{\mathcal{Y}_T p_{LR}}{12} (20p_{LS} - 21p_{LR} - 17 + 20p_{SR}) + \frac{\mathcal{Z}_T}{3} p_{LR}^2 \right) \end{aligned} \quad (18)$$

for geometry (b), if we ignore in both geometries terms of order θ_s^2 and $1/n$.

For a proof of this theorem see Section 6.

Remark 3.3.

- In the supporting online material you find a Mathematica-notebook for computing the exact values of the standard linkage disequilibrium measured by $\widehat{\sigma}_D^2$ without ignoring terms of order θ_s^2 and $1/n$.
- If the population evolves neutrally till the beginning of the sweep, Ohta and Kimura (1969) have shown, that

$$\begin{aligned} \mathcal{X}_T &= \frac{1}{4} \frac{\theta^2}{\theta + 1} \cdot \frac{5 + 2\theta + \rho_{LR}}{(1 + \theta)(3 + 2\theta + 2\rho_{LR})(5 + 2\theta + \rho_{LR}) - 2(3 + 2\theta)} - 4 \\ \mathcal{Y}_T &= \frac{\theta^2}{\theta + 1} \cdot \frac{1}{(1 + \theta)(3 + 2\theta + 2\rho_{LR})(5 + 2\theta + \rho_{LR}) - 2(3 + 2\theta)} \\ \mathcal{Z}_T &= \frac{1}{4} \frac{\theta^2}{\theta + 1} \cdot \frac{2\theta + \rho_{LR} + 5}{(1 + \theta)(3 + 2\theta + 2\rho_{LR})(5 + 2\theta + \rho_{LR}) - 2(3 + 2\theta)}, \end{aligned}$$

where $\theta := 4Nu$ is the neutral mutation rate. For a comparison of the theoretical results with simulations we assume that the population evolved neutrally till the beginning of the sweep.

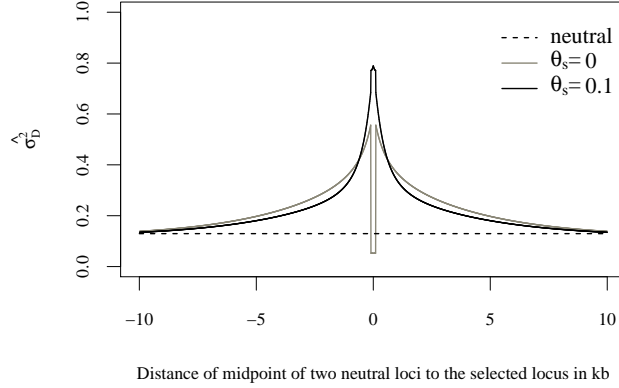


Figure 6: Theoretical values of $\hat{\sigma}_D^2$ in the neutral setting, for $\theta_s = 0$ and $\theta_s = 0.1$. The distance between the neutral loci is 0.2 kb, the selection strength $\alpha = 1000$, the population size $N = 10^6$, the recombination rate between the neutral loci $\rho_{LR} = 5$ and the neutral mutation rate $\theta = 0.005$.

- See Figure 6 for a plot of the theoretical values of $\hat{\sigma}_D^2$ for different values of θ_s . Here we assumed as well neutral evolution till the beginning of the sweep.
- Note, that for $\theta_s = 0$ we obtain $\hat{\sigma}_D^2$ for a hard sweep, compare (Pfaffelhuber et al., 2008).

4 Simulations

We simulated sequence samples with the new program msms (for ms mit Selektion (German: with selection)) of Greg Ewing, see (Ewing and Hermisson, 2010) to compare our theoretical linkage disequilibrium values with linkage disequilibrium values obtained from simulated genealogies assuming neutral evolution till the beginning of the sweep. The program msms generates sequence samples for a single selected locus of a population reproducing according to the Wright-Fisher-model with the possibility of recurrent mutation to the beneficial allele. The frequency of the beneficial allele is simulated stochastically conditioned on fixation. In Section 2 we argued, that the star-like genealogies approximates the Moran model genealogies well. For large population sizes the Moran model and Wright-Fisher model deliver similar genealogies, if the parameters are appropriately scaled. So instead of comparing the theoretical results with results obtained from Moran model simulations, we can check the theoretical results against results obtained from Wright-Fisher model simulations.

We consider a 5-kb stretch of DNA in a sample of $n = 20$ taken at time of fixation of the beneficial allele. We divide the stretch into 50 bins, each of length 0.1kb and measure LD between SNPs of two different bins averaged over 10^4 draws. Figure 7 shows the results for a recurrent mutation rate $\theta_s = 0.1$ and $\theta_s = 0.5$, respectively. The neutral mutation rate $\theta = 0.005$, the recombination rate $\rho = 0.025$ for two neighboring loci, the distance between neighboring neutral loci L and R is 200bp, the selection strength $\alpha = 1000$ and the population size $N = 10^6$ in both plots. These parameter values are realistic for example for *Drosophila melanogaster* samples.

As we see in Figure 7 there is a good fit between simulated and theoretical values of $\hat{\sigma}_D^2$. For small sample sizes the extended star-like genealogies approximate the simulated Wright-Fisher-genealogies well. The linkage disequilibrium is for theoretical and simulated values high, if $\theta_s \neq 0$

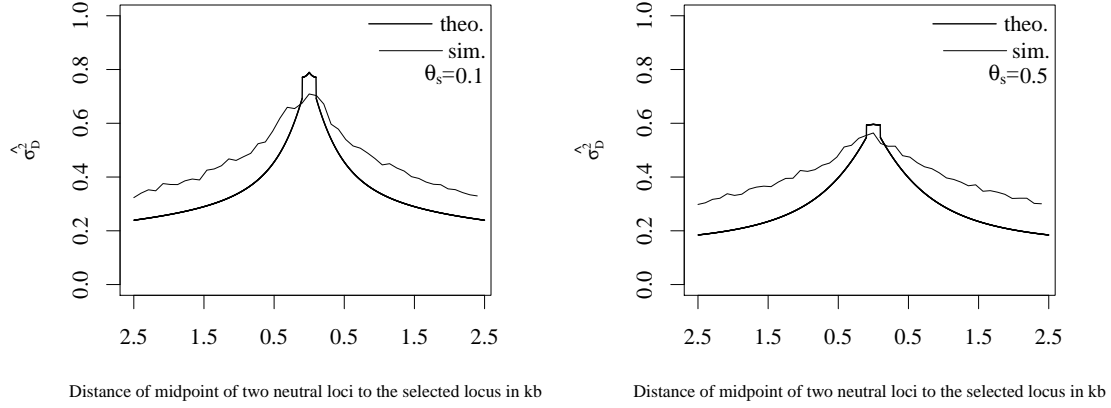


Figure 7: Left figure: Plot of $\hat{\sigma}_D^2$ for a neutral mutation rate $\theta = 0.005$, recombination rate $\rho = 0.025$, selection strength $\alpha = 1000$, recurrent mutation rate to the beneficial allele $\theta_s = 0.1$, a distance of 200 bp between the two neutral loci and a DNA-stretch of length 5 kb based on 10^4 draws. Right figure: Plot of $\hat{\sigma}_D^2$ with the same parameters as in the left figure except for the recurrent mutation rate to the beneficial allele $\theta_s = 0.5$.

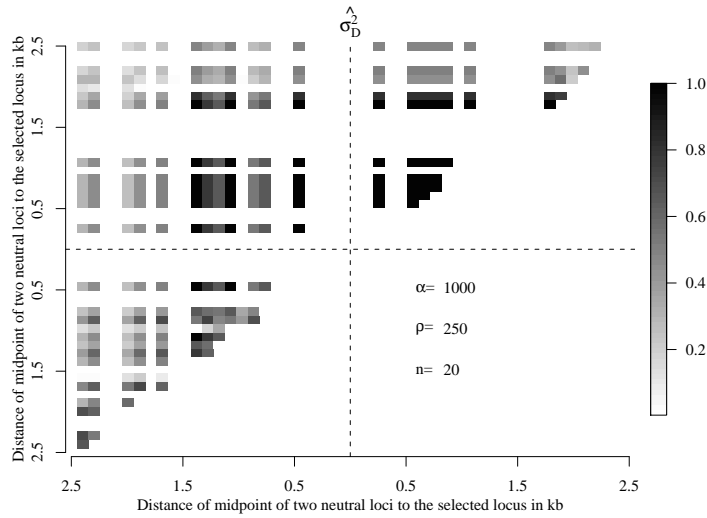


Figure 8: The full linkage disequilibrium spectrum for a single sample of a soft sweep with two founders with respect to the beneficial locus.

and the distance to the selected locus is small, independently of the geometry of the considered neutral loci. Due to recombination linkage disequilibrium decreases with increasing distance to the selected locus.

The differences between the theoretical and simulated values of $\widehat{\sigma_D^2}$ are due to the approximation of the genealogy. The approach has three effects on LD.

First, a star-like genealogy assumes independent recombinants. But of course in the simulated genealogies may also occur coalescence events before recombination events, in particular may arise early recombinants (see (Durrett and Schweinsberg, 2004) or for slightly different models (Etheridge et al., 2006)). For geometry (a) it is important, that recombinants with offspring lead to less “independent” variation, which can be seen in higher LD values of the simulated data. For geometry (b) early recombinants become noticeable, because they produce patterns similar to soft sweep patterns: Recombinants spread through the populations act as additive founders of the sweep. For this reason our approximation of the genealogy assumes less founders of the sweep than the simulated genealogies have. Therefore the LD-patterns of simulated data should look like the LD-pattern of the theoretical values with a slightly higher θ_s value. Higher θ_s values produce in geometry (b) less $\widehat{\sigma_D^2}$, compare the pictures in Figure 7. This effect becomes negligible for increasing θ_s . On the one hand the fixation of the beneficial allele gets faster, on the other hand for intermediate and high values of θ also extended star-like genealogies assume in average more than two founders. By measuring linkage disequilibrium one can distinguish well between the existence of one or two founders of the sweep, but not between the existence of three or four founders.

Second, the star-like approximation of the genealogy is in general longer than the simulated genealogy, since the beneficial allele spreads faster through the population, if the lines are dependent. Therefore, more recombination events are assumed to fall on the theoretical genealogies than on the simulated genealogies. For loci in a small neighborhood of the selected locus this means, that more SNPs can be found for the star-like genealogy due to recombination. Third, SNPs of simulated data are noisier, they may exist also due to neutral mutation during the sweep. Both effects can in geometry (a) be recognized by higher theoretical LD values in a small neighborhood of the selected locus. But at a certain point, the effect turns over: More recombination brings more “independent” variation into the sample: The theoretical values of LD in geometry (a) lie below the simulated values.

Often one is interested in the case of a single sample. We simulated a single sample of a soft sweep with two founders and computed linkage disequilibrium with that data. The result is plotted in Figure 8. In that case the pattern is very clear. In the neighborhood of the selected locus, high linkage disequilibrium can be found independent of the geometry of the loci. However, such clear patterns cannot be expected in general, even if the sweep has two founders. It is likely, that the number of offspring is not distributed equally between the founders. For example it may happen, that in a sample of 20 individuals with respect to the beneficial allele 2 individuals are offspring of one founder and the remaining 18 individuals are offspring of the second founder. For such unbalanced cases stochastic effects caused by recombination and mutation destroy the pattern easily.

5 Discussion

Soft sweeps have been introduced by Pennings and Hermisson in their series of papers (Hermisson and Pennings, 2005), (Pennings and Hermisson, 2006a), (Pennings and Hermisson, 2006b). They argued, that tests based on haplotype structure have high power to detect soft sweeps. Linkage disequilibrium is a test sensitive to haplotype structure. If allelic variants are tightly linked to a haplotype, LD is high for pairs of such alleles. We have seen, that linkage disequilibrium under a non vanishing recurrent mutation rate differs sufficiently from linkage disequilibrium under neutrality and hard sweeps, see Figure 6.

We computed $\widehat{\sigma_D^2}$ to understand the interplay of haplotype formation due to a soft sweep and

recombination. The main reason to compute $\widehat{\sigma_D^2}$ instead of $\mathbb{E}[r^2]$ is its mathematical manageability. However, former studies show (and the present study does that also), that also $\widehat{\sigma_D^2}$ measures what intuitively is understood under linkage disequilibrium and gives a possibility to distinguish between different population genetics scenarios.

When a soft sweep occurs, recombination breaks up the linkage of loci due to haplotype structure. Under hard sweeps recombination causes linkage of loci lying on one side of the selected locus. In Figure 6 theoretical values of $\widehat{\sigma_D^2}$ are plotted for different values of θ_s and under neutrality. The behavior can be explained roughly in the following manner:

For small values θ_s we see for both geometries high values of $\widehat{\sigma_D^2}$ in a small neighborhood of the selected locus decreasing with increasing distance to the selected locus. If θ_s is relatively small, Pennings and Hermisson (2006a) have shown, that soft sweeps are not very likely, most sweeps will be hard. LD of hard sweeps depends on recombination. Only recombination brings variation into the sample which is necessary to compute linkage disequilibrium.

After a hard sweep we can see the following pattern of LD due to recombination. Recombination between the L -locus and the S -locus includes for geometry (a) always a recombination between the R -locus and S -locus or between the L -locus and the R -locus, i.e. the L -locus recombines not independently of the R -locus. Therefore LD is high for geometry (a) for a hard sweep. In geometry (b) an LS -recombination event does not cause a SR -recombination event and vice versa. So with respect to recombination the L -locus is independent of the R -locus. Hence $\widehat{\sigma_D^2}$ is small. If the sample is not finite, $\widehat{\sigma_D^2}$ is even zero, see Remark 3.3.

If a soft sweep occurred, different founders of the sweep bring the variation into the sample - recombination is not necessary. If there are exactly two founders and there exist loci with two allelic variants, such that one allelic variant is carried by one haplotype and the other allele by the other haplotype, two of such loci are tightly linked, only recombination can break up this linkage. Therefore after a soft sweep with only a few number of founders LD is high in a small neighborhood of the selected locus, independent of the geometry. But the more founders the soft sweep has, the more variation is in the sample not linked to single founder. This reduces LD.

For r^2 we expect for very small values of θ_s patterns of LD similar to hard sweeps, because for very small values of θ_s soft sweeps are rare. But $\widehat{\sigma_D^2}$ shows even for very small values of θ_s high values in a small neighborhood of the selected locus. This comes from the fact, that small values of D^2 expected after a hard sweep in geometry (b) have a smaller effect on the numerator of $\widehat{\sigma_D^2}$ than higher values of D^2 expected after a soft sweep in geometry (b). An analogous statement holds for the denominator of $\widehat{\sigma_D^2}$.

For biological studies often the pattern of a single selective sweep is of interest. After a soft sweep we expect to find high LD of two neutral loci lying in a neighborhood of the selected locus, but almost neutral variation. It can be found haplotype structure, where each founder of the sweep gives rise to one haplotype. In each haplotype group a hard sweep occurred, i.e. almost no variation can be found, low LD for neutral loci lying on different sides of the selected locus and high LD for loci lying on the same site of the selected locus. In Figure 8 simulation results of a single sample of a soft sweep with two ancestors with respect to the selected locus are shown. As well as Tishkoff et al. (2007) found a comparable clear linkage disequilibrium pattern of a soft sweep in their studies of the human DNA when analyzing the human lactase persistence in African and European human populations.

An adaptation process may not only be initiated by mutation, but also through recurrent migration or from standing genetic variation during an environmental change. A two-island model with the beneficial allele fixed in one of the islands and migration from this island to the other coincides with our model for recurrent mutation. A more realistic model assumes, that the beneficial allele is not fixed in both islands and that the allelic frequencies q_L , q_R , etc. do not coincide on both islands. Unfortunately such (simple) modifications make the calculations in the proof of Theorem 3.2, especially of matrix A and B , quite complicated.

An improvement of the results could be made by approximating the genealogy not by a star-like approximation but by a marked Yule process with immigration. It has been shown by Hermisson

and Pfaffelhuber (2008), that the joint genealogy of the population is better approximated by these processes. However, explicit calculations become with this approximation complicated, since recombination is not independent along lines during the sweep.

6 Proof of Theorem 3.2

We proceed in five steps. The quantities $\mathcal{X}_t, \mathcal{Y}_t, \mathcal{Z}_t$ can be expressed in pairwise heterozygosities. In step 1 we will give this connection. In step 2 we show, how pairwise heterozygosities are transformed to sample heterozygosities. In step 3 and 4 we show, how pairwise heterozygosities at time $t = T$ are transformed to pairwise heterozygosities at time $t = 0$. In step 5 we collect everything together.

Step 1: Link between the pairwise heterozygosities f_t, g_t, h_t and $\mathcal{X}_t, \mathcal{Y}_t, \mathcal{Z}_t$

The quantities $\mathcal{X}_t, \mathcal{Y}_t, \mathcal{Z}_t$ can be expressed in terms of probabilities for pairwise heterozygosities.

Denote for this purpose by f_t the probability that two pairs heterozygous in both loci are linked, by g_t the probability, that exactly one pair of the two pairs is linked and the other pair is unlinked and by h_t the probability that both pairs are unlinked at time t . We can express these probabilities in terms of structured partitions: Let ℓ_1, ℓ_2 be two L -loci and r_1, r_2 be two R -loci taken from the population. Let $\xi_t = (\xi_t^B, \xi_t^b)$ be the genealogy of $\{\ell_1, \ell_2, r_1, r_2\}$ at time t , then

$$\begin{aligned} f_t &= P(\xi_t \text{ is heterozygous and } \xi_t^B \cup \xi_t^b = \{\{\ell_1, r_1\}, \{\ell_2, r_2\}\}), \\ g_t &= P(\xi_t \text{ is heterozygous and } \xi_t^B \cup \xi_t^b = \{\{\ell_1, r_1\}, \{\ell_2\}, \{r_2\}\}), \\ h_t &= P(\xi_t \text{ is heterozygous and } \xi_t^B \cup \xi_t^b = \{\{\ell_1\}, \{r_1\}, \{\ell_2\}, \{r_2\}\}) \end{aligned}$$

From an easy calculation (see for details also (Pfaffelhuber et al., 2008), Equation (A3)) it follows, that

$$\begin{pmatrix} \mathcal{X}_t \\ \mathcal{Y}_t \\ \mathcal{Z}_t \end{pmatrix} = \underbrace{\frac{1}{4} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 4 & -4 \\ 1 & -2 & 1 \end{pmatrix}}_{=: E} \begin{pmatrix} f_t \\ g_t \\ h_t \end{pmatrix}$$

Step 2: Link between pairwise heterozygosities f, g, h and sample heterozygosities $\hat{f}, \hat{g}, \hat{h}$

Denote by \hat{f}_t, \hat{g}_t and \hat{h}_t the corresponding sample probabilities, i.e. $\ell_1, \ell_2 \in \ell$ and $r_1, r_2 \in \mathbf{r}$. It is possible to pick the same individual twice in a sample. Therefore the following relationships hold:

$$\begin{aligned} \hat{f}_t &= \left(1 - \frac{1}{n}\right) f_t \\ \hat{g}_t &= \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) g_t + \left(1 - \frac{1}{n}\right) \frac{1}{n} f_t \\ \hat{h}_t &= \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \left(1 - \frac{3}{n}\right) h_t + \left(1 - \frac{1}{n}\right) \frac{4}{n} \left(1 - \frac{2}{n}\right) g_t + \left(1 - \frac{1}{n}\right) \frac{2}{n} \frac{1}{n} f_t. \end{aligned}$$

Denoting

$$F := I + \frac{1}{n} \begin{pmatrix} -1 & 0 & 0 \\ 1 & -3 & 0 \\ 0 & 4 & -6 \end{pmatrix} + \frac{1}{n^2} \begin{pmatrix} 0 & 0 & 0 \\ -1 & 2 & 0 \\ 2 & -12 & 11 \end{pmatrix} + \frac{1}{n^3} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -2 & 8 & -6 \end{pmatrix},$$

$$\text{this is equivalent to } \begin{pmatrix} \hat{f}_t \\ \hat{g}_t \\ \hat{h}_t \end{pmatrix} = F \begin{pmatrix} f_t \\ g_t \\ h_t \end{pmatrix}$$

For example, two linked pairs of one allele at the L - and one allele at the R -locus each taken at random (with replacement) from a sample are heterozygous, if we did not pick the same individual twice and the resulting two different lines are heterozygous at both loci.

In the next two steps we compute how to find f_0, g_0 and h_0 given f_T, g_T and h_T , respectively.

Step 3 : From $f_{T/2}, g_{T/2}$ and $h_{T/2}$ to f_0, g_0 and h_0 , respectively

For both geometries we have

$$\begin{pmatrix} f_0 \\ g_0 \\ h_0 \end{pmatrix} = C \begin{pmatrix} f_{T/2} \\ g_{T/2} \\ h_{T/2} \end{pmatrix}$$

with

$$C = \begin{pmatrix} p_{LR}^2 & 2p_{LR}(1-p_{LR}) & (1-p_{LR})^2 \\ 0 & p_{LR} & 1-p_{LR} \\ 0 & 0 & 1 \end{pmatrix}$$

Our model assumptions coincide with the model assumptions of Pfaffelhuber et al. (2008) in the time interval $[T, T/2]$, so that we obtain the same results here.

Step 4 : From f_T, g_T and h_T to $f_{T/2}, g_{T/2}$ and $h_{T/2}$, respectively

For this time step it is important to note, that it has to be paid attention not only on the two neutral loci, but also on the selected locus. We use Ewens sampling formula (see Equation 10) to compute the probabilities, if the ancestral lines of the pairs share with respect to the selected locus a common ancestors or different ancestors. With this we get the following relationships:

$$\begin{pmatrix} f_{T/2} \\ g_{T/2} \\ h_{T/2} \end{pmatrix} = A \begin{pmatrix} f_T \\ g_T \\ h_T \end{pmatrix} \quad \text{for geometry (a)}$$

and

$$\begin{pmatrix} f_{T/2} \\ g_{T/2} \\ h_{T/2} \end{pmatrix} = B \begin{pmatrix} f_T \\ g_T \\ h_T \end{pmatrix} \quad \text{for geometry (b)}$$

with matrix

$$A = (a_{ij})_{1 \leq i, j \leq 3}$$

given by

$$\begin{aligned}
a_{11} &= \frac{\theta_s}{\theta_s+1} p_{LR}^2 + \frac{1}{\theta_s+1} (1 - p_{SL}^2) p_{LR}^2 \\
a_{12} &= \frac{\theta_s}{\theta_s+1} 2p_{LR}(1 - p_{LR}) + \frac{1}{\theta_s+1} 2(1 - p_{SL}^2) p_{LR}(1 - p_{LR}) \\
a_{13} &= \frac{\theta_s}{\theta_s+1} (1 - p_{LR})^2 + \frac{1}{\theta_s+1} (1 - p_{SL}^2) (1 - p_{LR})^2 \\
a_{21} &= \frac{\theta_s}{(\theta_s+1)(\theta_s+2)} p_{SL} p_{SR} p_{LR} + \frac{2}{(\theta_s+1)(\theta_s+2)} (1 - p_{SL}) p_{SL}^2 p_{LR}^2 \\
a_{22} &= \frac{\theta_s^2}{(\theta_s+1)(\theta_s+2)} p_{LR} + \frac{\theta_s}{(\theta_s+1)(\theta_s+2)} (3p_{LR} - 3p_{SR} p_{SR}) + \\
&\quad + \frac{2}{(\theta_s+2)(\theta_s+1)} p_{LR} (1 - p_{SL}) (1 + p_{SL} - 4p_{SL} p_{SR} + 2p_{SL}^2) \\
a_{23} &= \frac{\theta_s^2}{(\theta_s+1)(\theta_s+2)} (1 - p_{LR}) + \\
&\quad + \frac{\theta_s}{(\theta_s+1)(\theta_s+2)} (1 - p_{LR}) (1 - p_{SL}^2 + 1 - p_{SL} p_{SR} + 1 - p_{SL} p_{SR}) + \\
&\quad + \frac{2}{(\theta_s+1)(\theta_s+2)} (1 - p_{SL}) (1 - p_{LR}) (1 + p_{SL} - 2p_{SL} p_{SR}) \\
a_{31} &= \frac{2\theta_s}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} p_{SL}^2 p_{SR}^2 \\
a_{32} &= \frac{4\theta_s^2}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} p_{SL} p_{SR} + \\
&\quad + \frac{2\theta_s}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} (p_{SR} p_{SL} (1 - p_{SR} p_{SL})) + \\
&\quad + \frac{4\theta_s}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} p_{SL} p_{SR} (2 - p_{SL} - p_{SR}) + \\
&\quad + \frac{3!}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} (4(1 - p_{SL}) (p_{SL} (1 - p_{SR}) p_{SR})) \\
a_{33} &= \frac{\theta_s^3}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} + \\
&\quad + \frac{\theta_s^2}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} (4(1 - p_{SL} p_{SR}) + (1 - p_{SL} p_{SL}) + (1 - p_{SR} p_{SR})) + \\
&\quad + \frac{4\theta_s}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} ((1 - p_{SR}) (1 - p_{SL}) (1 - p_{SR}) + (1 - p_{SR}) (1 - p_{SR}) p_{SL}) + \\
&\quad + \frac{4\theta_s}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} (2(1 - p_{SR}) (1 - p_{SL}) p_{SR} + (1 - p_{SL}) (1 - p_{SL}) p_{SR}) + \\
&\quad + \frac{4\theta_s}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} (2(1 - p_{SR}) p_{SL} (1 - p_{SL}) + (1 - p_{SL}) (1 - p_{SL}) (1 - p_{SR})) + \\
&\quad + \frac{\theta_s}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} (2(1 - p_{SL} p_{SR}) (1 - p_{SL} p_{SR}) + (1 - p_{SL} p_{SL}) (1 - p_{SR} p_{SR})) + \\
&\quad + \frac{6}{(\theta_s+1)(\theta_s+2)(\theta_s+3)} ((1 - p_{SL}) (1 - p_{SR}) (1 + p_{SL} + p_{SR} - 3p_{SL} p_{SR}))
\end{aligned}$$

and matrix

$$B = (b_{ij})_{1 \leq i, j \leq 3}$$

given by

$$\begin{aligned}
b_{11} &= \frac{\theta_s}{\theta_s + 1} p_{LS}^2 p_{SR}^2 \\
b_{12} &= \frac{2\theta_s}{\theta_s + 1} p_{SR} p_{LS} (1 - p_{LS} p_{SR}) + \frac{4}{\theta_s + 1} p_{LS} (1 - p_{LS}) p_{SR} (1 - p_{SR}) \\
b_{13} &= \frac{\theta_s}{\theta_s + 1} (1 - p_{LS} p_{SR})^2 + \\
&\quad + \frac{1}{\theta_s + 1} (1 - p_{LS}) (1 - p_{SR}) (2p_{LS} (1 - p_{SR}) + 2(1 - p_{LS}) p_{SR} + (1 - p_{LS}) (1 - p_{SR})) \\
b_{21} &= \frac{\theta_s}{(\theta_s + 1)(\theta_s + 2)} p_{LS}^2 p_{SR}^2 \\
b_{22} &= \frac{\theta_s^2}{(\theta_s + 1)(\theta_s + 2)} p_{LS} p_{SR} + \\
&\quad + \frac{\theta_s}{(\theta_s + 1)(\theta_s + 2)} (p_{LS} p_{SR} (1 - p_{LS} p_{SR}) + (2 - p_{SR} - p_{LS}) p_{SR} p_{LS} + p_{SR} p_{LS} (1 - p_{LS} p_{SR})) + \\
&\quad + \frac{8}{(\theta_s + 1)(\theta_s + 2)} p_{LS} (1 - p_{LS}) p_{SR} (1 - p_{SR}) \\
b_{23} &= \frac{\theta_s^2}{(\theta_s + 1)(\theta_s + 2)} (1 - p_{SR} p_{LS}) + \\
&\quad + \frac{\theta_s}{(\theta_s + 1)(\theta_s + 2)} ((1 - p_{LS})^2 (1 - p_{SR}) + 2p_{LS} (1 - p_{LS}) (1 - p_{SR})) + \\
&\quad + \frac{\theta_s}{(\theta_s + 1)(\theta_s + 2)} (p_{SR} (1 - p_{LS})^2 + (1 - p_{LS} p_{SR})^2 + (1 - p_{SR})^2 (1 - p_{LS})) + \\
&\quad + \frac{\theta_s}{(\theta_s + 1)(\theta_s + 2)} (2p_{SR} (1 - p_{SR}) (1 - p_{LS}) + p_{LS} (1 - p_{SR})^2) \\
&\quad + \frac{2}{(\theta_s + 1)(\theta_s + 2)} (1 - p_{LS}) (1 - p_{SR}) (2p_{LS} (1 - p_{SR})) \\
&\quad + \frac{2}{(\theta_s + 1)(\theta_s + 2)} (2(1 - p_{LS}) p_{SR} + (1 - p_{LS}) (1 - p_{SR})) \\
b_{31} &= a_{31} \quad \text{with } p_{LS} \text{ instead of } p_{SL} \\
b_{32} &= a_{32} \quad \text{with } p_{LS} \text{ instead of } p_{SL} \\
b_{33} &= a_{33} \quad \text{with } p_{LS} \text{ instead of } p_{SL}
\end{aligned}$$

To see the above equations, consider for example in geometry (a) the term a_{21} : In this case at time $T/2$ there are two pairs, which are heterozygous in both loci and exactly one of the pairs is linked. If the two pairs have two different ancestors with respect to the selected locus neither a SL -recombination nor a SR -recombination must happen for the unlinked pair, nor a LR -recombination event to the linked pair. Therefore the probability to stay linked also at the beginning of the sweep is $\frac{\theta_s}{(\theta_s+1)(\theta_s+2)} p_{SL} p_{SR} p_{LR}$ using Ewens sampling formula. If the two pairs have a single ancestor, the linked pair has to change backgrounds, i.e. a SL -recombination event has to take place. Therefore we obtain in this case the probability $\frac{2}{(\theta_s+1)(\theta_s+2)} (1 - p_{SL}) p_{SL}^2 p_{LR}^2$. The sum of these two probabilities gives a_{21} . The other terms can be explained in an analogous manner.

Step 5 : Collecting all together

We have

$$(\widehat{\mathcal{X}}_0, \widehat{\mathcal{Y}}_0, \widehat{\mathcal{Z}}_0)^T = E \cdot F \cdot C \cdot A \cdot E^{-1} (\mathcal{X}_T, \mathcal{Y}_T, \mathcal{Z}_T)^T$$

for geometry (a) and

$$(\widehat{\mathcal{X}}_0, \widehat{\mathcal{Y}}_0, \widehat{\mathcal{Z}}_0)^T = E \cdot F \cdot C \cdot B \cdot E^{-1}(\mathcal{X}_T, \mathcal{Y}_T, \mathcal{Z}_T)^T$$

for geometry (b).

With this we can compute $\widehat{\sigma}_D^2 = \widehat{\mathcal{Z}}_0 / \widehat{\mathcal{X}}_0$ (recalling Equation (14)).

A calculation with Mathematica gives Equations (15)-(18), if terms of order θ_s^2 and $1/n$ are ignored. \square

Acknowledgement

I am grateful to Peter Pfaffelhuber for many fruitful discussions. Many thanks to Greg Ewing for providing and helping me with msms and to Franz Baumdicker and Joachim Hermisson for helpful comments on the manuscript. I acknowledge support from the DFG Forschergruppe 1078 "Natural selection in structured populations".

References

- Barton, N., Etheridge, A., and Sturm, A. (2004). Coalescence in a random background. *Ann. of Appl. Probab.*, 14, no. 2:754–785.
- Durrett, R. and Schweinsberg, J. (2004). Approximating Selective Sweeps. *Theo. Pop. Biol.*, 66(2):129–138.
- Etheridge, A., Pfaffelhuber, P., and Wakolbinger, A. (2006). An approximate sampling formula under genetic hitchhiking. *Ann. Appl. Probab.*, 16:685–729.
- Ewing, G. and Hermisson, J. (2010). Msms: A coalescent simulation program including recombination, demographic structure, and selection at a single locus. *Bioinformatics*, *submitted*.
- Hermisson, J. and Pennings, P. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352.
- Hermisson, J. and Pfaffelhuber, P. (2008). The pattern of genetic hitchhiking under recurrent mutation. *Elec. J. Probab.*, 13(68):2069–2106.
- Karasov, T., Messer, P. W., and Petrov, D. A. (2010). Evidence that adaptation in drosophila is not limited by mutation at single sites. *PLoS Genet*, 6(6).
- Kim, Y. and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167:1513–1524.
- Kurtz, T. G. (1971). Limit theorems for sequences of jump markov processes approximation ordinary differential processes. *J. Appl. Prob.*, 8:344–356.
- Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favorable gene. *Genetic Research*, 23:23–35.
- McVean, G. A. (2007). The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175:1395–1406.
- Nair, S., Nash, D., Sudimack, D., Jaidee, A., Barends, M., Uhlemann, A., Krishna, S., Nosten, F., and Anderson, T. (2007). Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Mol. Biol. Evol.*, 24:562–573.
- Ohta, T. and Kimura, M. (1969). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, 63(1):229–238.

- Pennings, P. and Hermisson, J. (2006a). Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.*, 23(5):1076–1084.
- Pennings, P. and Hermisson, J. (2006b). Soft Sweeps III - The signature of positive selection from recurrent mutation. *PLoS Genetics*, 2(e186).
- Pfaffelhuber, P., Lehnert, A., and Stephan, W. (2008). Linkage disequilibrium under genetic hitchhiking in finite populations. *Genetics*, 179:527–537.
- Pritchard, J. K., Pickrell, J. K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*, 20(4):208–215.
- Scheinfeldt, L. B., Biswas, S., Madeoy, J., Connelly, C. F., Schadt, E. E., and Akey, J. M. (2009). Population genomic analysis of *alms1* in humans reveals a surprisingly complex evolutionary history. *Mol Biol Evol*, 26(6):1357–1367.
- Schlenke, T. A. and Begun, D. J. (2005). Linkage disequilibrium and recent selection at three immunity receptor loci in *drosophila simulans*. *Genetics*, 169(4):2013–2022.
- Song, Y. and Song, J. (2007). Analytic computation of the expectation of the linkage disequilibrium coefficient r^2 . *Theo. Pop. Biol.*, 71:49–60.
- Stephan, W., Song, Y. S., and Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, 172(4):2647–2663.
- Tishkoff, S., Reed, F., Ranciaro, A., Voight, B., Babbitt, C., Silverman, J., Powell, K., Mortensen, H., Hirbo, J., Osman, M., Ibrahim, M., Omar, S., Lema, G., Nyambo, T., Gori, J., Bumpstead, S., Pritchard, J., Wray, G., and Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.*, 39:31–40.